Issues in the Planning and Design of a Bilingual (English–Northern Sotho) Explanatory Dictionary for Industrial Electronics

Elsabé Taljard, *Department of African Languages, University of Pretoria, Pretoria, Republic of South Africa (elsabe.taljard@up.ac.za),*

Rachélle Gauton, *Department of African Languages, University of Pretoria, Pretoria, Republic of South Africa (rachelle.gauton@up.ac.za)*
and
Liam A. Gauton, *NamITech Pty. Ltd. and Department of African Languages, University of Pretoria, Pretoria, Republic of South Africa (liam.gauton@namitech.com)*

Abstract: The focus of this article is the planning and design of a bilingual, explanatory dictionary for industrial electronics with a clearly delimited and very specific target user in mind. Since the number of lemmas to be treated in the dictionary is limited to 500, special care must be taken to select those lemmas that are relevant for both the purpose of the dictionary and the needs of the target user. It is indicated that the main consideration in the planning of the envisaged dictionary is user-friendliness, as dictated by the intended target users. In this article, a novel approach to the semi-automatic selection of lemmas for inclusion in an LSP dictionary is described. The procedure that is used for the extraction of definitional information from an electronic corpus is also explained.

Keywords: LSP LEXICOGRAPHY, DICTIONARY PLANNING, LEMMA SELECTION, SEMI-AUTOMATIC TERM EXTRACTION, DEFINITIONAL INFORMATION, INDUSTRIAL ELECTRONICS, CORPUS-BASED LEXICOGRAPHY

Opsomming: Kwessies by die beplanning en ontwerp van 'n tweetalige (Engels–Noord-Sotho) verklarende woordeboek vir industriële elektronika. Die fokus van hierdie artikel is die beplanning en ontwerp van 'n tweetalige, verklarende woordeboek vir industriële elektronika met 'n duidelik afgebakende en baie spesifieke teikengebruiker in gedagte. Aangesien die getal lemmas vir behandeling in die woordeboek tot 500 beperk is, moet besondere sorg gedra word dat daardie lemmas gekies word wat beantwoord aan die doel van die woordeboek én die behoeftes van die teikengebruiker. Daar word uitgewys dat die hoofoorweging in die beplanning van die beoogde woordeboek gebruikersvriendelikheid is, soos bepaal deur die bestemde teikengebruikers. In hierdie artikel word 'n nuwe benadering tot die semi-outomatiese keuse van lemmas vir insluiting in 'n vakwoordeboek beskryf. Die prosedure wat vir die onttrekking van definisiële inligting uit 'n elektroniese korpus gebruik word, word ook verduidelik.

Lexikos 17 (AFRILEX-reeks/series 17: 2007): 152-169

Sleutelwoorde: LEKSIKOGRAFIE VIR VAKWOORDEBOEKE, WOORDEBOEKBEPLANNING, LEMMASELEKSIE, SEMI-OUTOMATIESE TERMONTTREKKING, DEFINISIËLE INLIGTING, INDUSTRIËLE ELEKTRONIKA, KORPUS-GEBASEERDE LEKSIKOGRAFIE

1. Introduction

The research upon which this article is based, forms part of a larger project entitled Language, Educational Effectiveness and Economic Outcomes, thus LE^{3}O, done under the auspices of the Centre for Research in the Politics of Language (CentRePoL) at the University of Pretoria. The main aim of this project as stated by Webb (2005: 5) is to contribute towards addressing poverty and inequality in the distribution of wealth through the development of the vocational literacy of workers. This is done by "researching the role of language in vocational education and training, proposing ways in which the fundamental role of language in vocational training and practice can be recognised and corrected, and proposing specific policies regarding the role of language in Further Education and Training in South Africa". One of the subprojects is the introduction of Northern Sotho as medium of instruction in a course in industrial electronics (IE) taught at two colleges for Further Education and Training (FET) in the Greater Metropolitan Area of Tshwane. An offshoot of this project is the compilation of a bilingual (English–Northern Sotho) explanatory dictionary for industrial electronics (BEDIE). Due to financial and other constraints the number of terms to be treated in the dictionary is restricted to approximately 500. Due to this rather severe restriction, particular attention must be paid to the proper lexicographic and terminological planning of this (rather rudimentary) dictionary so that the maximum amount of relevant information can be imparted to the user in the most effective way. The planning of this dictionary therefore forms the focus of this article, with specific reference to the selection of the lemma list and the semi-automatic extraction of definitional information from an electronic English IE corpus. The actual compilation of the dictionary will take place in two phases. The first phase will consist of the selection of the English terms to be included in the dictionary, followed by the writing of terminological definitions for the selected terms. During the second phase, the terms and their definitions will be translated into Northern Sotho. The translation of the selected terms and their definitions from the original English source material into Northern Sotho is an undertaking generating its own pitfalls and challenges and does not fall within the ambit of the current investigation. (Cf. Gauton et al. (forthcoming) for an exposition of the difficulties inherent in this type of translation).

2. Intended purpose and function of the dictionary

Gouws and Prinsloo (2005: 13) point out that the purpose of a dictionary is determined by, inter alia, its typological nature and its intended target user

group. Typologically speaking, the envisaged dictionary will be a fully bilingual (English, Northern Sotho), unidirectional (English → Northern Sotho) explanatory dictionary. Taking the specific circumstances of the target user group into account, the dictionary will play an important role as a learning aid. The intention is to include the dictionary as an addendum of the textbook used in the instruction of industrial electronics so as to allow immediate access to the information provided in the dictionary. This would offer the users the opportunity of accessing definitions of basic industrial electronics terms in their home language, and would also provide them with Northern Sotho equivalents of key terms. The BEDIE is therefore primarily intended to assist learners with the decoding of texts written in a language foreign to them, i.e. English.

3. Intended target users

According to Bergenholtz and Tarp (1995: 20, 21), three crucial aspects need to be taken into account when the profile of the intended target user is drawn up, i.e. native-language competence, foreign-language competence and encyclopaedic knowledge. The intended target users of the BEDIE are learners aged 15–21 years, who are mother-tongue speakers of Northern Sotho, but who have hitherto received their formal education primarily through the medium of English, which is for most of them a second or even third language. Due to the fact that Northern Sotho has to a large extent not been used as a medium of instruction during their formal schooling, it is assumed that their academic proficiency in Northern Sotho is rather below the expected level, particularly within a technical field such as industrial electronics. This is confirmed by the preliminary results of an attitude test carried out by the LE^{3}O team at the two colleges involved in this project, as reported by Webb (2005: 121): 41.7% of learners who indicated that their home language is Northern Sotho, responded by saying that they cannot speak Northern Sotho well; 60.8% indicated that they cannot read Northern Sotho well, whereas 70% indicated that they are not adequately proficient in writing Northern Sotho. It can therefore be concluded that the native-language competence of the intended target user is below standard. With regard to their proficiency in English, i.e. foreign-language competence, proficiency tests revealed that their proficiency in English is also way below par, as is evident from examples extracted from written work done by the students during this test. Compare a few randomly selected examples of written material produced by learners during the English Language Proficiency Test as cited in Webb (2005: 12) and shown in example (1).

(1) — When we approaches the steps … there was so many people try to get their class … rember it was a serious injured … school help by paying everything to him … there is no job where you cold find one races only.

— I hope my situation will be requested soon … my friend miss to step on the stairs and it was hurt badly.

— He doesn't help the customer well. When you give the bus drive the bus you must already teach him the ruel of the road and keep the customer good. And they must obey the ruel of the customer; I will happy with the manager her us.

It is interesting to note the discrepancy between learners' perception of their own competence and their actual proficiency in English — according to Webb (2005: 119), 92% of respondents reported that they spoke English well/very well, 96.8% indicated that they could read it well/very well, 90.7% were of the opinion that they could write English well/very well, whereas 99.1% indicated that their comprehension of English is good/very good. This is in stark contrast to the actual English usage, as reflected in the examples above.

With regard to their encyclopaedic knowledge, it must be kept in mind that learners entering an FET college have had little, if any, exposure to a subject such as industrial electronics. Secondly, taking available assessment results into consideration, it must be concluded that even after training the intended target user has an average to low level of encyclopaedic knowledge and can therefore be regarded as a layperson. Compare in this regard Table 1 showing the results obtained for industrial electronics during the first three years of training at the two colleges, as cited by Webb (2005: 50).

Table 1: Assessment results, trimester 2, 2002

	Average %			Total average %
	N1 (Grade 10)	N2 (Grade 11)	N3 (Grade 12)	
College A	43	43	43	43
College B	75	65	44	61

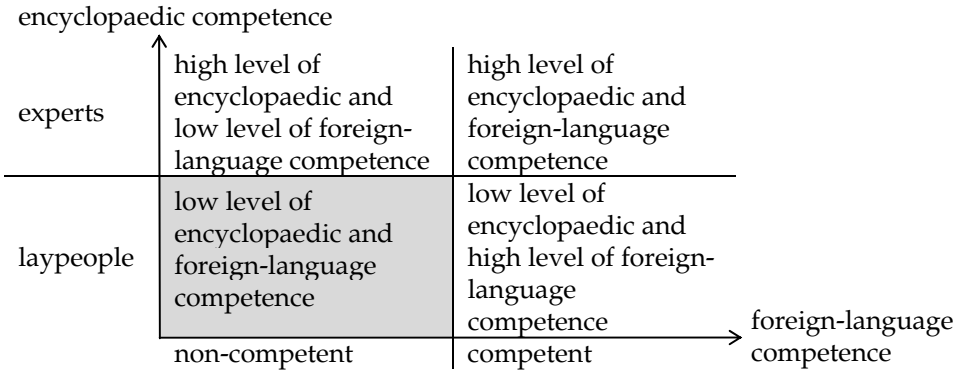
In terms of the four main types in user profile as distinguished by Bergenholtz and Tarp (1995: 21), the user profile for this particular project is indicated in Figure 1.

As Bergenholtz and Tarp (1995: 21) point out, it needs to be kept in mind that the transition between the different types is fluid.

Taking the broader South African situation with regard to the culture of dictionary use into account, it can furthermore be assumed that the target users have had very little, if any, exposure to the culture of dictionary use. According to Gouws and Prinsloo (2005: 42), the majority of South Africans find themselves in a "pre-dictionary culture environment". This is especially valid for speakers of the South African Bantu languages. Atkins (as quoted by Gouws and Prinsloo (2005: 42)) states that speakers of these languages "have not in their formative years had access to dictionaries of the richness and complexity of those currently available for European languages. They have not had the chance to internalize the structure and objectives of a good dictionary". If user-

friendliness is to be one of the primary considerations in the compilation of the BEDIE, cognisance needs to be taken of the almost complete lack of a dictionary culture and the resulting absence of dictionary consultation skills.

Figure 1: User profile: IE dictionary



4. Planning of the dictionary

The team members involved in the planning of the dictionary deemed it extremely important that the envisaged dictionary should be planned according to sound terminological and lexicographic principles. Taking both the profile of intended target-users and the purpose of the dictionary into account, it was decided that an approach that is user-friendly in the extreme, is the only appropriate one. This has direct implications for both the micro- and macrostructure of the intended dictionary. With regard to microstructure, the value of a rather simplified structure needs to be weighed up against the need to provide the maximum amount of encyclopaedic information. The maximum number of data categories that constitutes the dictionary article is five, including the lemma itself. Compare example (2).

(2) <lemma> negative type
<definition> piece of semiconductor doped with impurities, which enables it to donate electrons
<symbol> N-type
<cross-reference address> ALSO SEE: positive type
<synonym> (none)

The ordering typology is strictly alphabetical, rather than thematic as is often the case in LSP dictionaries. However, it has to be kept in mind that one of the functions of a good LSP dictionary is to reveal conceptual relationships existing between different terms. Bergenholtz and Tarp (1995: 199) point out that the alphabetical ordering principle precludes the illustration of conceptual relation-

ships, since concepts occur out of context in an arbitrary order. The user is therefore not given an overview of the conceptual structure of the subject field in question. Although this is a valid point, it needs to be kept in mind that all related concepts do not necessarily share the same headword. A thematic ordering would therefore go some way in revealing conceptual relationships, but does not offer the ultimate solution to the problem. On the other hand, an alphabetic arrangement is practical, fast and familiar to the user and, taking the profile of the target user of the BEDIE into consideration, probably the appropriate ordering principle. However, the exposure of conceptual relationships cannot be sacrificed for the sake of simplicity, therefore alternative strategies need to be employed. For the purpose of this dictionary, mainly two techniques are used. In the first instance, external cross-referencing is employed to explicitly indicate to the user that specific concepts are related. According to Gouws and Prinsloo (2005: 179), an external cross-reference address can either be located elsewhere in the central list, or in a separate text outside the central list. Both these cross-reference types are utilized in the BEDIE, although the second type is aimed not so much at revealing conceptual relationships, but rather at providing additional information from an external source. Compare the articles of *series circuit* and *parallel circuit* in (3) below as an example of cross-referencing to another article in the central lemma list.

(3) **series circuit** *Circuit* in which the components are connected end to end so that the *current* has only one path to follow through the circuit. **ALSO SEE:** *parallel circuit*

parallel circuit *Circuit* consisting of two *resistors*, connected side by side so that there is more than one path through which *current* can flow. **ALSO SEE:** *series circuit*

An explicit cross-reference marker **ALSO SEE** at the end of both articles refers the user to the related concept. By reading both definitions, the target user can conceptualize the nature of the relationship existing between the two concepts. Secondly, by identifying the correct superordinate concept for all related subordinate concepts, the nature of the relationship between the superordinate and its subordinates is revealed, albeit in a more implicit manner. Compare example (4) in this regard, where a logical relationship exists between the superordinate concept 'metallic element in Group 11 of the Periodic Table (a transition element)' and the two subordinate concepts 'silver' and 'copper'.

(4) **silver** A silver-white metallic element in Group 11 of the Periodic Table (a transition element), with *atomic number* 47. It conducts *current* and heat very well, even better than *copper* does. SYMBOL: Ag. **SEE** PERIODIC TABLE p. iv

copper A reddish metallic element in Group 11 of the Periodic Table (a

transition element), with *atomic number* 29. It is very much used in electrical work because it conducts *electricity* so well. However, *gold* and *silver* conduct electricity better than copper does. SYMBOL: Cu. **SEE** PERIODIC TABLE p. iv

To further enhance the user-friendliness of the dictionary, all terms used in any given definition that are treated in the main lemma list of the dictionary, are printed in italics. This convention would of course have to be explained in the user's guide that will form part of the front matter text. The team members also decided to make each terminological definition as extensive as possible, providing the user with more than just a basic paraphrase of meaning. This would compensate to a certain extent for the low level of encyclopaedic competence of the target user. Compare in this regard the definition of the term *current flow* in example (5).

(5) **current flow** Movement of free *electrons* through a particular material. While the *atoms* of some materials (the *conductors*) give up electrons from their outer orbits freely, other materials hold on to these outer electrons much more tightly. The number of free electrons available is the factor which determines the ease with which *current* will flow through a material. A material with few free electrons will only pass current unwillingly. In other words, it offers opposition to current flow. This opposition to current flow is called *resistance*.

An expanded data distribution structure is envisaged, with the outer text containing amongst others a user's guide, a list of commonly used abbreviations and symbols, frequently used formulae, the Periodic Table as well as a list of academic words such as *define*, *compare, identify, illustrate, motivate,* etc., necessitated by the low level of English proficiency of the intended target users. Where applicable, the user will be referred to the outer texts by means of an explicit cross-reference marker. In (4) above, the marker **SEE** PERIODIC TABLE (with an indication of the exact page number) in small capitals is used.[1]

5. Selection of the lemmas to be treated in the dictionary

According to Bergenholtz and Tarp (1995: 98), metalexicographical literature has thus far paid scant attention to the selection of lemmas for LSP dictionaries. They cite doubt on the part of the metalexicographer about the possibility of theory development as a possible reason for this theoretical constraint. They nevertheless suggest a number of fundamental methodological approaches for the selection of lemmas for inclusion in an LSP dictionary, stating that "quality in practical lexicography includes meticulous, goal-oriented selection of lemmata". The selection process described below has exactly that in mind and represents a novel approach to lemma selection for LSP dictionaries, going

beyond the frequency approach, which seems to have become standard practice for LGP lexicography.

In keeping with developments in the international arena regarding the compilation of LSP dictionaries, it was decided that a corpus-based approach would be followed for this project. The advantages of such an approach have often been stated, one of them being the wealth of encyclopaedic and linguistic information the corpus provides. As Shreve (2001: 773) remarks, the text in which a technical term occurs is an important source of information — not only on its usage, meaning and appropriateness, but also on the relationship it has with other terms. When terms co-occur in a text, conceptual relationships are established. Shreve even suggests that the larger conceptual structure of a special subject field cannot exist unless it has been established by or extracted from a corpus of texts. The proposed procedure for the compilation of the BEDIE has been designed to mine and to maximally utilise any information provided by the corpus. For this particular project, the utilisation of the corpus revolves around two main issues: firstly, the selection of the appropriate terms to be entered into the central lemma list, and secondly, the extraction of definitional information that can be used for the writing of the terminological definitions.

5.1 Compilation of a special-purpose corpus on industrial electronics (IE corpus)

The special-purpose corpus used for the compilation of the BEDIE consists of three textbooks and a separate glossary of terms that form the basis of the curriculum taught at the two FET colleges in the Greater Metropolitan Area of Tshwane referred to earlier. These sources are the following:

— Van Deventer, D.J. *Industrial Electronics N1*. 2000. Cape Town: Maskew Miller Longman.

— Kraft, J. *Industrial Electronics N2*. 2000. Cape Town: Maskew Miller Longman.

— Kraft, J. *Industrial Electronics N3*. 2000. Cape Town: Maskew Miller Longman.

— Kraft, J. *Glossary. Electrical Technology*. 2004. Unpublished glossary compiled by J. Kraft for use in the teaching of Industrial Electronics at the Pretoria West Campus of the Tshwane South College for Further Education and Training.

As the new Curriculum Statement for Industrial Electronics in the FET Band was not yet available when the corpus was being compiled, the compilers were compelled to base the BEDIE on the most recent textbooks used in the teaching

of the curricula at the two colleges, i.e. the IE textbooks for N1, N2 and N3 published in 2000.

As a first step, a small special-purpose corpus on industrial electronics was compiled in which each of the four sources mentioned above was treated as a subcorpus. This was done by scanning the material using *OmniPage* software, then storing the electronic version in ordinary .txt (text) format. This was done after the necessary permission had been obtained from the authors to reproduce the text in electronic format. The sizes of the various subcorpora within the IE corpus in running words (so-called 'tokens') are shown in Table 2.

Table 2: Subcorpora and sources

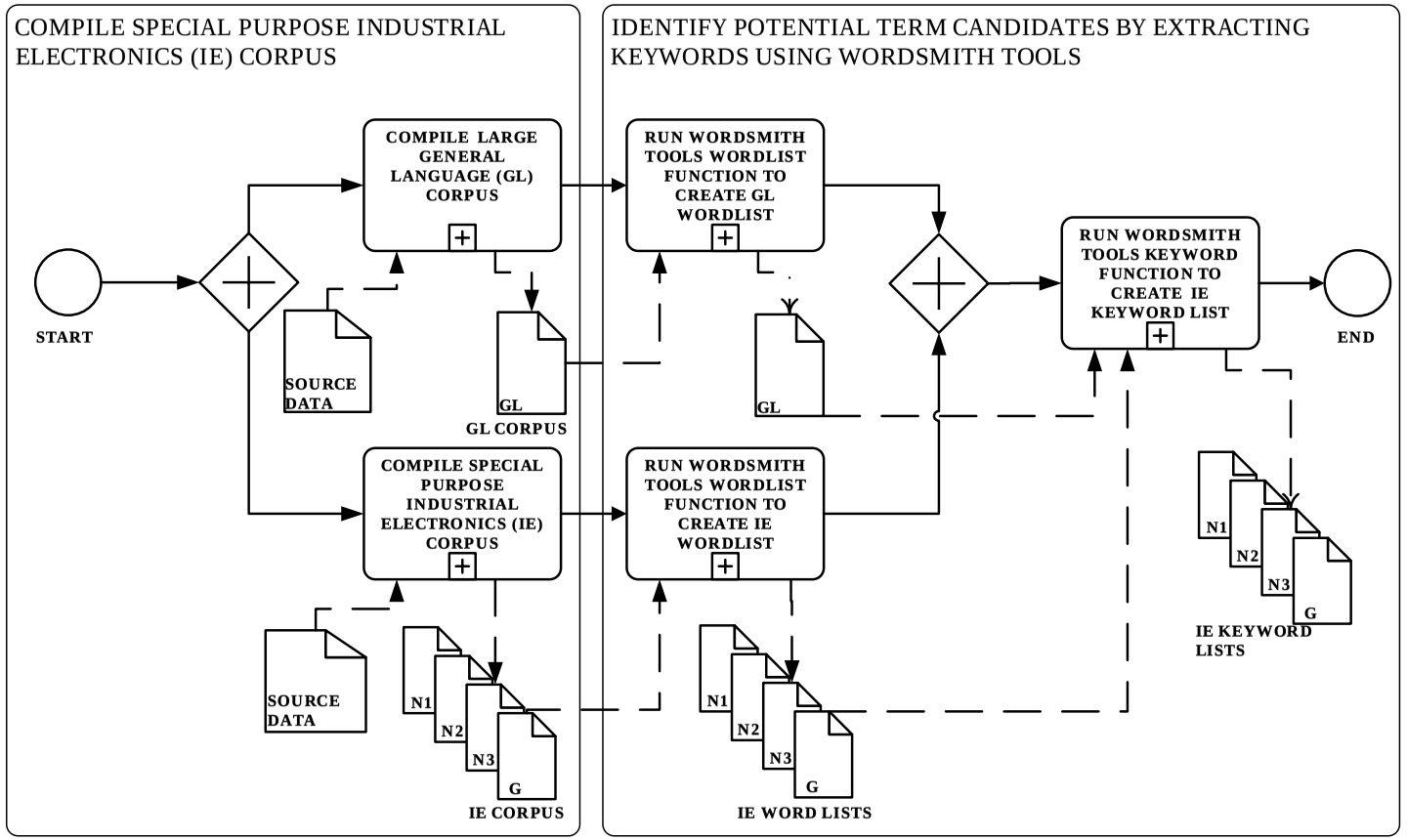
	Source	**Size in running words ('tokens')**
Subcorpus 1	N1	29 590
Subcorpus 2	N2	23 569
Subcorpus 3	N3	33 699
Subcorpus 4	Glossary	4 699
	TOTAL	**91 557**

5.2 Semi-automatic extraction of terms

In order to identify and semi-automatically extract the appropriate 500 terms to be treated in the dictionary, the corpus query software *WordSmith Tools* was used, in particular its WordList and KeyWords functions. In order to identify term candidates for possible inclusion in the BEDIE, the WordList function is applied first in order to compile a word list according to frequency of occurrence for each of the four special-field subcorpora *and* for a larger, general-language reference corpus. As a reference corpus, the University of Pretoria English Internet Corpus (PEIC) was used. This electronic corpus consists of 12.5 million English running words (tokens), culled from the Internet by Rachélle Gauton. The KeyWords tool is then used to locate and identify keywords in the four special-field subcorpora — the keywords being those words that characterise each of these texts. In order to identify such keywords, the word lists excerpted from each of the small special-purpose corpora (IE subcorpora 1–4) are compared with those of the large general corpus (PEIC). Words with an unusually high frequency in the small specialized corpora are regarded as 'key' and therefore as potential term candidates. This procedure is represented in Diagram 1.

The assumption that underlies the use of the KeyWords tool for the purpose of identifying appropriate terms for an LSP dictionary, is that words that appear with an outstanding frequency in a small special-purpose corpus, will probably be the most important or key *terms* in that corpus. It should be noted however, that the KeyWords list inevitably contains some 'noise', i.e. items that are obviously not terms. Some manual intervention is therefore needed to re-

Diagram 1: Procedure for the identification of keywords, i.e. potential term candidates



move these items from the list of term candidates. At this stage, the input of a special-field expert is of great value, since it is not always possible for non-specialists to decide whether a specific item is indeed to be regarded as a term in the specific subject field. For this particular project, the assistance of an expert in industrial electronics attached to the physics section of the Foundation Year Programme at the University of Pretoria was sought.[2]

By comparing each of the four IE subcorpora with the large PEIC corpus one-, two-, three-, four- and five-word keywords were subsequently identified for *each* of the subcorpora, representing the study years N1, N2 and N3 as well as the separate glossary of terms (G). (See again Diagram 1 in this regard.)

These keyword lists were then further analysed (through being filtered, ranked and selected) by making use of the open-source database management system *MySQL™*. Roughly speaking, each of the N1, N2 and N3 KeyWords lists represented a priority list of words/terms from each of three syllabus years that build upon the knowledge base of the learner. The working assumption was that there would be repetition from each preceding year's data in the following year's data. First, the *Wordsmith Tools* KeyWords output for each syllabus year was imported into its own database table. An *SQL* query to extract unique terms from the first year's data (i.e. N1) was written and the result set was stored in a new table used to build a consolidated list of unique terms. From the second year's data (i.e. N2), an *SQL* query was constructed to extract unique terms that did not already exist in the unique terms table and the result set was then added to the unique terms table. This process was repeated for the third year's data (N3) as well as for the separate glossary (G).

Once the consolidated list was available in its own database table, it was a simple matter to extract a list of the top 500 terms based on the (high) keyness value of the words as determined by the *WordSmith Tools* KeyWords function. (As stated earlier, words with a high (positive) keyness value are those words with an unusually outstanding (high) frequency when compared with the words in a large (general language) reference corpus). The words/terms in this list were then roughly lemmatized in order not to count e.g. singular and plural forms of the same word as two separate lemmas/headwords. This procedure can be represented visually in Diagram 2.

Subsequently, an attempt was made to validate the 500 identified term candidates by comparing them with keywords extracted from a 58 990 running words (tokens) IE corpus culled from the Internet by Rachélle Gauton (in consultation with Philip Pare). This corpus, although dealing with the field of industrial electronics in general, was compiled in such a way that it also focused on IE texts aimed at learners, although not necessarily South African learners. Again by using *SQL*, the top 500 term candidates list from the special-purpose corpus based on the curriculum taught at the two FET colleges was compared with a list of possible term candidates from the generally available IE Internet corpus to ensure the validity of the terms. On comparing these two lists, it was found that a total of 333 term candidates appear on both lists. A list

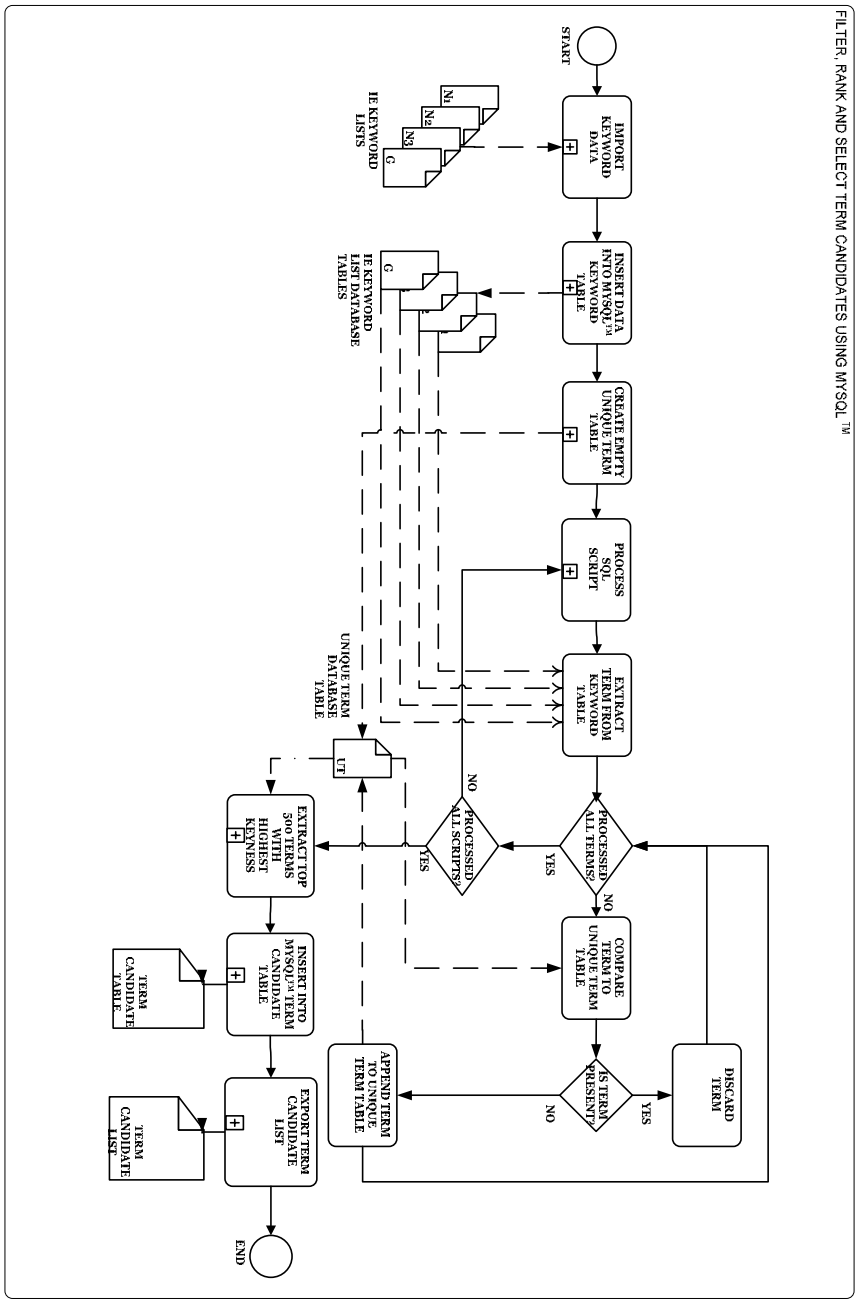


Diagram 2: Procedure for filtering, ranking and selecting keywords, i.e. potential term candidates, using *MySQL ™*

containing the remaining terms that appear in the top 500 term candidates list culled from the special-purpose IE corpus, but *not* in the generally available IE Internet corpus, was then presented to the subject specialist for comment. This was done to ascertain the acceptability or not (as the case may be) of these terms for inclusion in the LSP dictionary. This was essentially a cross-check procedure to make sure that terms are not identified as candidates for inclusion in the dictionary should they, for example, not be in use any more and/or incorrect. From the list of term candidates presented to him, the subject specialist selected 147 for inclusion in the BEDIE, resulting in a total of 480 single and multiword term candidates — a number that is close to the original target of 500. This procedure is represented in Diagram 3.

However, what had not been taken into account up to this point, was the fact that the definitions themselves generate more terms for treatment in the dictionary. It is a generally accepted principle that terms used in terminological definitions should also be defined in an LSP dictionary. The definition of the term *capacitance* in (6) serves as an example.

(6) **capacitance** Property of an object which opposes a change in *voltage*, e.g. the ability of a *capacitor* to store *energy* in an electric field. Capacitance depends upon the distance between the ***plates*** of a capacitor, as well as on the area of the plates. The larger the plate area, the bigger the *charge* that can be stored. Capacitance is measured in ***farads***

Six terms, printed in italics, are used in the definition. Four of these appear on the 480 core term list, but the two terms *plate(s)* and *farad(s)* do not. The recommendation of the subject expert was that these two terms needed to be lemmatised and treated. Thus the definition of the term *capacitance* generated two new terms for inclusion in the lemma list of the dictionary.

A further consideration that comes into play is that of completing paradigms of closely related concepts. The concepts 'proton', 'electron' and 'neutron' are a case in point. The terms *proton* and *electron* both appear in the 480 core list of term candidates, but not the term *neutron*. However, such a close conceptual relationship exists between these three concepts, that it would be a serious oversight not also to lemmatize and fully treat the term *neutron*.

Consequently, an analysis was made of the definitions of the 100 terms with the highest keyness values in the 480 core term list (sorted according to keyness). It was found that a total of 28 extra terms, i.e. terms that do not appear on the core term list, were generated by the definitions of these terms, either because they were used in the definitions or because they formed part of a larger paradigm. At this stage, it is not yet clear whether the 28% increase is a constant that should be taken into consideration by terminologists and LSP lexicographers when selecting terms to be included in an LSP dictionary for which the number of terms to be treated is restricted. It is also not clear if the same kind of increase would present itself when moving down the keyness list,

Diagram 3: Procedure for validating the top 500 term candidates extracted from the special-purpose IE corpus (based on the curriculum taught at the two FET colleges) using *MySQL ™*

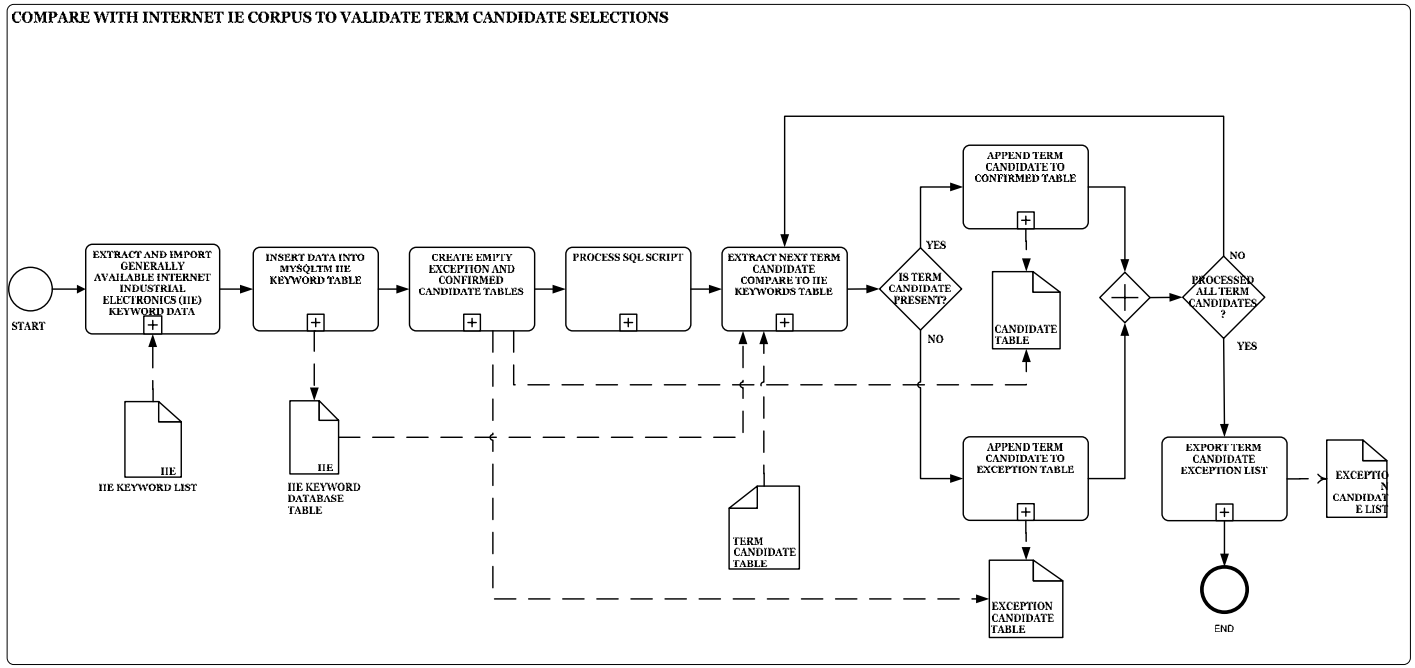
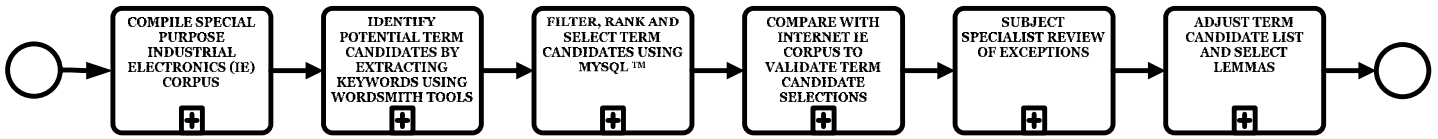


Diagram 4: Procedure followed for the selection of lemmas to be treated in the BEDIE



in other words whether the definitions of the second 100 key terms would generate the same percentage of extra terms. Further investigation is needed before such an increase can be factored in.

In summary, the overall procedure followed in selecting the lemmas to be treated in the dictionary, can be visually represented as in Diagram 4.

6. Semi-automatic extraction of definitional information and compilation of terminological definitions

The 333 term candidates which appear as KeyWords in both the IE corpus based on the curriculum taught at the two FET colleges and the IE Internet corpus aimed at learners, together with the 147 terms (from the top 500 list of term candidates extracted from the special-purpose IE corpus based on the curriculum taught at the two FET colleges) that were added on the recommendation of the subject expert, are regarded as representing the core vocabulary of the IE texts used to build the corpora. As these 480 identified term candidates have been sorted according to their keyness values, definitional information can then be extracted from the text corpora by entering each of the term candidates in order of keyness as a search node in the concordancing function KeyWords in Context (KWIC) of *WordSmith Tools*. The information thus retrieved forms the basis of the terminological definitions, starting at the top 100 key terms. The basic assumption underlying this method is that authors of technical texts very often provide definitional information in these texts. The technique used to semi-automatically retrieve definitional information from electronic texts is described in Pearson (1998: 103). She indicates that either lexical or syntactic devices signal the presence of definitional information in texts. A typical syntactic device is the use of the copula, whereas typical lexical markers are expressions such as *is/are called*, *is/are known as*, etc. Since the IE corpus that is used for the compilation of the BEDIE is untagged, mainly lexical markers were relied upon for the identification of definitional information. Even so, since the IE corpus is relatively small (91 557 tokens), manual scanning of all KWIC lines generated for a specific search word is not a huge undertaking, thus the search for definitional information is not restricted to the identification of information marked by lexical items. Compare the example in (7) below, representing a sample of KWIC lines in which the term *electromagnet* is the search node.

(7) Sample of KWIC lines with *electromagnet* as search node

10. Such a device is known as an	**electromagnet**	and is used extensively in the
Magnetism remaining in the core of an	**electromagnet**	after the coil current is removed.
correct polarity to prevent breakdown.	**electromagnet**	A coil of wire usually wound on a
A relay makes use of an	**electromagnet.**	When current flows through the coil
that is tripped or activated by use of an	**electromagnet.**	magnetic coil: Spiral of a conductor

By making use of the Grow facility of *WordSmith Tools*, more context surrounding the search node can be revealed, as shown in (8).

(8) More context surrounding the search node *electromagnet* as revealed by the Grow function

As iron provides a better path (higher permeability) for the lines of force than air, the strength of the magnetism is much greater as shown in figure 6.10. Such a device is known as an **electromagnet** and is used extensively in the manufacture of electronic equipment such as relays, doorbells, buzzers and circuit breakers.

---

remanence Amount a material remains magnetized after the magnetizing force has been removed. residual magnetism Magnetism remaining in the core of an **electromagnet** after the coil current is removed. resistance Symbolized "R" and measured in ohms.

The oxide acts as the dielectric for the capacitor. Electrolytic capacitors are polarized and so must be connected in correct polarity to prevent breakdown. **electromagnet** A coil of wire usually wound on a soft iron or steel core. When current is passed through the coil a magnetic field is generated. The core provides an easy path for the magnetic lines of force. This concentrates the field in the core.

In order to use the coil as a controlled electromagnet, it should be wound on a core that retains little of its magnetism. A relay makes use of an **electromagnet**. When current flows through the coil, the iron core is magnetised.

---

magnet Body that can be used to attract or repel magnetic materials. magnetic circuit breaker Circuit breaker that is tripped or activated by use of an **electromagnet**. magnetic coil Spiral of a conductor which is called an electromagnet. magnetic core Material that exists in the center of the magnetic coil to either physically support the windings (non-magnetic material) or to concentrate the magnetic flux (magnetic material).

Based on the definitional information retrieved from the KWIC lines, the definition of the term *electromagnet* in (9) was formulated.

(9) **electromagnet** Magnet which consists of a *coil* of wire that is usually wound on a soft *iron* or *steel* core. When *current* is passed through the coil a *magnetic field* is generated. The core provides an easy path for the *magnetic lines of force*. Electromagnets are widely used in the manufacture of electronic equipment such as doorbells, buzzers, relays and *circuit* breakers.

After the definitions based on information retrieved from the KWIC lines have been formulated, these are submitted to the subject field expert for final checking of contents. At this stage of the compilation process, the input of the subject-field expert is essential, since the lexicographer — not being an expert in industrial electronics — has no way of judging whether the salient features of a particular concept had indeed been thrown up by the corpus. Where necessary, definitions are then revised. Compare example (10) in this regard, where (a) is the definition compiled by the lexicographer, based on definitional information extracted from the corpus, and (b) is the definition as revised by the expert.

(10) (a) **tunnel diode** *Semi-conductor* device that will exhibit a negative *resistance* between the values of 0.2 *volts* and 0.4 volts when *forward-biased*. It has a low peak tunnel *current* and is used as a low-voltage *rectifier* biased in the reverse direction.

(b) **tunnel diode** *Semi-conductor* device that has a negative *resistance* between the values of about 0.2 *volts* and 0.4 volts when *forward-biased*. It is usually used as a high-speed switch and in high-frequency *oscillator circuits*. These are *circuits* which are used to produce high-frequency *alternating currents*. It was invented by Leo Esaki in 1958.

In cases where no definitional information can be retrieved from the corpus, definitions are supplied by the subject-field expert.

One of the main advantages of the procedure described above is the fact that very little claim is laid to the knowledge of the terminologist/LSP lexicographer of the particular subject field. The retrieval of definitional information from the corpus is to a large extent a mechanical process. Furthermore, the demands made on the time and input of the subject-field expert are also lessened — something which might make experts more inclined to take part in such projects.

7. Translation of terms and definitions

After final checking by the subject-field expert, the list of lemmas and their definitions are handed over to the translator for rendering them into Northern Sotho. As indicated in the introduction, a discussion of the translation process does not fall within the ambit of this article and will therefore not be addressed here.

8. Conclusion

In this article, a possible approach to the compilation of an LSP dictionary was discussed by presenting the planning and design of the BEDIE as a case in point.

The main contribution this article makes to the field of (corpus) lexicography, and specifically to the discipline of (corpus) LSP lexicography, is that it details a highly effective and functional approach towards the selection of lemmas and subsequent extraction of definitional information for an LSP dictionary in the most labour-efficient manner (i.e. computer aided as opposed to manual) by using various functionalities of (general purpose) software such as *WordSmith Tools* and *MySQL ™*.

Notes

1. The page number cited in this example is a fictional one, added for illustrative purposes.
2. We would like to express our thanks to Phillip Pare, who unstintingly shared his time and expertise with us. Without his input, this project would not have been possible.

Bibliography

Bergenholtz, H. and S. Tarp (Eds.). 1995. *Manual of Specialised Lexicography: The Preparation of Specialised Dictionaries.* Amsterdam/Philadelphia: John Benjamins.

Gauton, R., E. Taljard, T.A. Mabasa and L.F. Netshitomboni. Forthcoming. Translating Technical (LSP) Texts into the Official South African Languages: A Corpus-based Investigation of Translators' Strategies. (To be submitted to *Language Matters*).

Gouws, R.H. and D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography*. Stellenbosch: SUN PReSS.

Kraft, J. 2000. *Industrial Electronics N2*. Cape Town: Maskew Miller Longman.

Kraft, J. 2000. *Industrial Electronics N3*. Cape Town: Maskew Miller Longman.

Kraft, J. 2004. *Glossary. Electrical Technology*. Unpublished glossary used in the teaching of Industrial Electronics at the Pretoria West Campus of the Tshwane South College for Further Education and Training.

MySQL™ Version 5.0.1. Available from http://www.mysql.com.

OmniPage. Available from http://www.nuance.com.

Pearson, J. 1998. *Terms in Context.* Amsterdam: John Benjamins.

Scott, M. 1999. *WordSmith Tools Version 3*. Oxford: Oxford University Press. Also available from http://www.lexically.net/wordsmith/index.html.

Shreve, G.M. 2001. Terminological Aspects of Text Production. Wright, Sue Ellen and Gerhard Budin (Eds.). 2001. *Handbook of Terminology Management. Volume 2. Application-Oriented Terminology Management*: 772-787. Amsterdam: John Benjamins.

Van Deventer, D.J. 2000. *Industrial Electronics N1*. Cape Town: Maskew Miller Longman.

Webb, V.N. 2005. *Report on the Project Language, Educational Effectiveness and Economic Outcomes.* Submitted to the Swiss Development Agency. Pretoria: University of Pretoria, CentRePoL.